Mathematical Description of the Multivariate Stratified Sampling Strategy

Ameir Shaa

February 11, 2024

This document provides a detailed mathematical framework for implementing a 3D histogram-based sampling strategy, aimed at preserving the joint distribution of three-dimensional spatial data represented by coordinates (X, Y, Z).

1 Framework

Given a dataset $\mathcal{D} = \{(x_i, y_i, z_i) \mid i = 1, \ldots, N\}$, where each (x_i, y_i, z_i) represents a point in 3D space and N is the total number of points, our goal is to sample a subset $\mathcal{S} \subset \mathcal{D}$ such that $\mathcal{S}$ maintains the statistical properties of $\mathcal{D}$ across (X, Y, Z) dimensions.

2 Creating a 3D Histogram

The first step involves partitioning the 3D space into bins that collectively form a 3D histogram. This process is mathematically represented as follows:

$$
\begin{aligned}
&\text{Let } B_x, B_y, B_z \text{ be the set of bins for the } X, Y, Z \text{ dimensions respectively.}\\
&\quad B_x = \{b_1^x, b_2^x, \ldots, b_{n_x}^x\},\\
&\quad B_y = \{b_1^y, b_2^y, \ldots, b_{n_y}^y\},\\
&\quad B_z = \{b_1^z, b_2^z, \ldots, b_{n_z}^z\},\\
&\text{where}\\
&\quad b_j^x = [x_{\min}^j, x_{\max}^j),\\
&\quad b_k^y = [y_{\min}^k, y_{\max}^k),\\
&\quad b_l^z = [z_{\min}^l, z_{\max}^l),\\
&\quad \text{for } j = 1, \ldots, n_x; k = 1, \ldots, n_y; l = 1, \ldots, n_z.
\end{aligned}
$$

Each bin b_m^d in dimension $d \in \{X, Y, Z\}$ with index m defines a range $[\min^m, \max^m)$ that categorizes points based on their d-coordinate.

3 Stratification and Sampling

The dataset is stratified based on the 3D histogram, with each bin representing a stratum. Sampling is performed within each stratum to achieve a representative subset:

1. For each bin (b_j^x, b_k^y, b_l^z) in the 3D histogram, select a random subset of points $\mathcal{S}_{jkl} \subset \mathcal{D} \mid |\mathcal{S}_{jkl}| = \lceil 0.1 \cdot |\mathcal{D}_{jkl}| \rceil$
2. with $\mathcal{D}_{jkl} = \{(x_i, y_i, z_i) \in \mathcal{D} \mid x_i \in b_j^x, y_i \in b_k^y, z_i \in b_l^z\}$,
3. and $|\mathcal{D}_{jkl}|$ denotes the number of points in $\mathcal{D}$ that fall within the bin (b_j^x, b_k^y, b_l^z).

4 Aggregating Sampled Data

The final sampled subset $\mathcal{S}$ is obtained by aggregating the sampled subsets from all bins:

$$\mathcal{S} = \bigcup_{j=1}^{n_x} \bigcup_{k=1}^{n_y} \bigcup_{l=1}^{n_z} \mathcal{S}_{jkl}.$$

5 Conclusion

This 3D histogram-based sampling strategy ensures that the sampled subset $\mathcal{S}$ is representative of the original dataset $\mathcal{D}$, preserving the joint distribution of (X, Y, Z) while reducing the size of the dataset.